| | |
|---|---|
| Title | Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks |
| Authors | Walsh, Aaron M.;Crispie, Fiona;Daari, Kareem;O'Sullivan, Orla;Martin, Jennifer C.;Arthur, Cornelius T.;Claesson, Marcus J.;Scott, Karen P.;Cotter, Paul D. |
| Publication date | 2017-06-16 |
| Original Citation | Walsh, A. M., Crispie, F., Daari, K., O'Sullivan, O., Martin, J. C., Arthur, C. T., Claesson, M. J., Scott, K. P. and Cotter, P. D. (2017) 'Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks', Applied and Environmental Microbiology. In Press, doi: 10.1128/aem.01144-17 |
| Type of publication | Article (peer-reviewed) |
| Link to publisher's version | 10.1128/aem.01144-17 |
| Rights | © 2017 American Society for Microbiology |
| Download date | 2023-05-05 00:07:42 |
| Item downloaded from | http://hdl.handle.net/10468/4188 |

# Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks

Aaron M. Walsh[a,b,c], Fiona Crispie[a,b], Kareem Daari[d], Orla O'Sullivan[a,b], Jennifer C. Martin[d], Cornelius T. Arthur[e], Marcus J. Claesson[b,c], Karen P. Scott[d], Paul D. Cotter[a,b,*].

[a]Teagasc Food Research Centre, Moorepark, Fermoy, Co. Cork, Ireland

[b]APC Microbiome Institute, University College Cork, Co. Cork, Ireland

[c]Microbiology Department, University College Cork, Co. Cork, Ireland

[d]Rowett Institute, University of Aberdeen, Aberdeen, Scotland, UK. AB25 2ZD

[e]Animal Research Institute, Accra, Ghana

[*]Corresponding author.

Address correspondence to:

Paul D. Cotter,
Principal Research Officer,
Teagasc Food Research Centre,
Moorepark, Fermoy, Co. Cork,
Ireland.
Email: paul.cotter@teagasc.ie
Telephone: +353 (0)25 42694

## Abstract

The rapid detection of pathogenic strains in food products is essential for the prevention of disease outbreaks. It has already been demonstrated that whole metagenome shotgun sequencing can be used to detect pathogens in food but, until recently, strain-level detection of pathogens has relied on whole metagenome assembly, which is a computationally demanding process. Here, we demonstrate that three short read alignment-based methods, MetaMLST, PanPhlAn, and StrainPhlAn, can accurately, and rapidly, identify pathogenic strains in spinach metagenomes which were intentionally spiked with Shiga toxin-producing *Escherichia coli* in a previous study. Subsequently, we employ the methods, in combination with other metagenomics approaches, to assess the safety of nunu, a traditional Ghanaian fermented milk product which is produced by the spontaneous fermentation of raw cow milk. We show that nunu samples are frequently contaminated with bacteria associated with the bovine gut, and worryingly, we detect putatively pathogenic *E. coli* and *Klebsiella pneumoniae* strains in a subset of nunu samples. Ultimately, our work establishes that short read alignment-based bioinformatics approaches are suitable food safety tools, and we describe a real-life example of their utilisation.


## Importance

Foodborne pathogens are responsible for millions of illnesses, annually. Here, we demonstrate that short read alignment-based bioinformatics tools can accurately, and rapidly, detect pathogenic strains in food products from shotgun metagenomics data. The methods used here are considerably faster than both traditional culturing methods and alternative bioinformatics approaches that rely on metagenome assembly, and thus they can potentially be used for more high-throughput food safety testing. Overall, our results suggest that whole metagenome sequencing can be used as a practical food safety tool to prevent diseases or link outbreaks to specific food products.

## Introduction

In recent years, high-throughput sequencing (HTS) has become an important tool in food microbiology (1). HTS enables in-depth characterisation of food-related microbial isolates,

51 *via* whole genome sequencing (WGS), and it facilitates culture-independent analysis of
52 mixed microbial communities in foods, *via* metagenomic sequencing.

53 WGS has provided invaluable insights into the genetics of starter cultures (2, 3), and it is
54 routinely used in epidemiology to identify outbreak-associated foodborne pathogens isolated
55 from clinical samples, by comparing the single nucleotide polymorphism (SNP) profiles of
56 outbreak strain genomes versus non-outbreak strain genomes (4-6). Metagenomic sequencing
57 enables the elucidation of the roles of microorganisms during food production (7-9), and it
58 can be used to track microorganisms of interest through the food production chain, as
59 illustrated by Yang *et al.* (10), who used whole metagenome shotgun sequencing to track
60 pathogenic species in the beef production chain. Indeed, metagenomic sequencing can be
61 used to detect pathogens in foods to monitor outbreaks of foodborne illnesses (11), but few
62 studies have done so, because of the limited taxonomic resolution achievable using these
63 methods. Typically, 16S rRNA gene sequencing provides genus-level taxonomic resolution
64 (12), and although sub-genus-level classification is achievable using species-classifiers (13)
65 or oligotyping (14, 15), these methods cannot accurately discriminate between strains.
66 Similarly, metagenome sequence classification tools usually provide species-level resolution
67 (16). However, strain-level resolution is necessary for the accurate identification of pathogens
68 in food products (17). Leonard *et al.* successfully achieved strain-level resolution of Shiga
69 toxin producing *Escherichia coli* strains in spinach samples using metagenome shotgun
70 sequencing (18). However, the bioinformatics methods used in that study were based on
71 metagenome assembly, which is a computationally demanding process (19, 20), and thus
72 alternative strain-level identification methods are needed.

73 Since 2016, several short read alignment based software applications, including MetaMLST
74 (20), StrainPhlAn (21), and PanPhlAn (19), have been released that can achieve strain-level
75 characterisation of microorganisms from metagenome shotgun sequencing data. All three
76 applications are considerably faster than metagenome assembly based methods. To date,
77 these programs have not been employed to detect pathogens in food products, but there is
78 strong evidence to suggest that they have considerable potential for this purpose: MetaMLST
79 accurately predicted that the strain responsible for the 2011 German *E. coli* outbreak
80 belonged to *E. coli* ST678 (20), and similarly, PanPhlAn accurately predicted that the strain
81 was a Shiga toxin producer (19), based on the analysis of the gut metagenomes of infected
82 patients (22). StrainPhlAn has so far not been used for epidemiological purposes, but a recent

study demonstrated that it can be used to predict the phylogenetic relatedness of bacterial strains from different samples (21).

MetaMLST aligns sequencing reads against a housekeeping gene database to identify sequence types present in metagenomic samples based on multilocus sequence typing (MLST). The MetaMLST database contains all currently known sequence types, but it can be updated as required to include newly identified sequence types. MetaMLST does not require any prior knowledge of the microbial composition of sample and it can simultaneously detect different species' sequence types. PanPhlAn aligns sequencing reads against a species pangenome database, constructed from reference genomes, to functionally characterise strains present in metagenomic samples. PanPhlAn allows the user to generate customisable pangenome databases for any species. StrainPhlAn extracts species specific marker genes from sequencing reads and it aligns the markers against reference genomes to identify the strains present in metagenomic samples. StrainPhlAn requires output from MetaPhlAn2, and both programs use the same database.

In this study, we describe the characterisation of nunu, a traditional Ghanaian fermented milk product (FMP), at the genus, species, and strain-levels, using a combination of 16S rRNA gene sequencing and whole metagenome shotgun sequencing. Nunu is produced by the spontaneous fermentation of raw cow milk in calabashes or plastic or metal containers under ambient conditions, and it is usually consumed after 24-36 hours (23). At present, little is known about nunu's microbiology, relative to other FMPs, like kefir or yoghurt (24). Previously, a number of potentially pathogenic bacteria, including *Enterobacter*, *Escherichia* and *Klebsiella*, were detected in nunu by culture based methods (25). Here, we carry out the first culture-independent analysis of a number of nunu samples. In addition to detecting the presence of a variety of lactic acid bacteria (LAB) typical of fermented dairy products, MetaMLST, PanPhlAn and StrainPhlAn all indicated the presence of pathogenic *E. coli* and *Klebsiella pneumoniae* in a subset of the samples. We also demonstrate that these tools can accurately predict the presence of pathogenic strains in foods by testing them on food metagenomes which were spiked with Shiga toxin producing *E. coli*. Ultimately, our work establishes that short read alignment based methods can be used for the detection of pathogens in foods.

## Results

**16S rRNA gene sequencing of nunu samples**

Nunu samples were collected from producers with hygiene practice training (n=5) and producers without hygiene practice training (n=5), respectively. 16S rRNA gene sequencing analysis revealed that there were no significant differences in the alpha-diversity of nunu samples from trained or untrained producers (Figure S1a), although there was a clear separation in the beta-diversity of the two groups (Figure S1b).

The 16S rRNA data was also analysed to determine bacterial composition (Figure 1a). At the family level, all of the samples were dominated by Lactobacillales, and at the genus-level, most samples were dominated by *Streptococcus*, although the sample 1t2am was dominated by *Lactococcus*. *Enterococcus* was detected in 4/10 samples (1 trained and 3 untrained) at ≥3% relative abundance, and it was highest in the sample 2u6am, where it was present at 19% relative abundance. In addition, *Staphylococcus* was detected in all 10 samples, although its abundance was ≤1% in each case. The detection of staphylococci was consistent with a corresponding culture-dependent analysis of the samples (supplemental material). Importantly, Enterobacteriales were also prevalent. *Enterobacter* was detected in 9/10 samples (4 samples from trained producers and 5 from untrained producers) at ≥1% relative abundance, and it was highest in the sample 2u8am, where it was present at 23% relative abundance. *Escherichia-Shigella* was detected in 8/10 samples (4 trained and 4 untrained) at ≥1% relative abundance, and it was highest in the sample 1t7am, where it was present at 17% relative abundance; this finding was again consistent with culture-dependent analysis of the samples (supplemental material).

The Kruskal-Wallis test indicated that there were significant differences in the relative abundances of *Macrococcus* (p=0.01), which was higher in samples from trained producers, and *Streptococcus* (p=0.02), which was higher in samples from untrained producers (Figure 1b). No other genera were significantly different.

**Species-level compositional analysis of nunu samples as revealed by shotgun sequencing**

MetaPhlAn2-based analysis of shotgun metagenomic data provided results that were generally consistent with those derived from amplicon sequencing. 11 species accounted for >90% of the microbial composition of every sample (Figure 2). At the species-level, most samples were dominated by *Streptococcus infantarius*, although sample 1t2am was

dominated by *Lactococcus lactis. Enterococcus faecium* was detected in 4/10 samples (2 trained and 2 untrained) at ≥1% relative abundance, and it was highest in the sample 1t2am, where it was present at 22% relative abundance. High abundances of Enterobacteriales were again apparent. *Enterobacter cloacae* were detected in the sample 1t8am, where it was present at 1% relative abundance. *Escherichia coli* was detected in 2/10 samples (2 trained) at ≥7% relative abundance, and it was highest in 1t7am, where it was present at 13% relative abundance. *Klebsiella pneumoniae* was detected in 7/10 samples (4 trained and 3 untrained) at ≥3% relative abundance, and it was highest in 1t8am, where it was present at 71% relative abundance. In contrast, *Klebsiella* was not detected by amplicon sequencing, and this discrepancy might be due to similarities in the 16S rRNA genes from these genera(42).

The Kruskal-Wallis test indicated that there were significant differences in the relative abundances of *Macrococcus caseolyticus* (p=0.01), which was higher in samples from trained producers, and *Streptococcus infantarius* (p=0.01), which was higher in samples from untrained producers (Figure S2). No other species were significantly different.

**Investigation of the functional potential of the nunu microbiota**

SUPER-FOCUS was used to provide an overview of the functional potential of the nunu metagenome. As expected, a significant proportion of the metagenome was assigned to housekeeping functions like carbohydrate metabolism, nucleic acid metabolism, and protein metabolism (Figure 3). However, SUPER-FOCUS also detected high levels of functions associated with horizontal gene transfer and virulence in nunu. The level 1 subsystem "Phages, Prophages, Transposable elements" was present at ≥1% average relative abundance in both groups, although it was significantly higher in nunu samples from trained producers (p=0.047). Similarly, the level 1 subsystem "Virulence" was present at ≥3.5% average relative abundance in both groups.

HUMAnN2 was used to provide more comprehensive insights into the functional potential of the nunu metagenome. Unsurprisingly, the 25 most abundant genetic pathways were associated with carbohydrate metabolism, nucleic acid metabolism, and protein metabolism (Figure 4a). MDS analysis of all the normalised HUMAnN2 pathway abundances suggested that there were differences in the overall functional potential of the groups (Figure S3), and we detected significant differences in the relative abundances of some individual pathways (Table S1). Notably, we observed that histidine degradation pathways were higher in trained

178 samples (p=0.047) (Figure 4c). Furthermore, histidine decarboxylase genes were only
179 detected in trained samples. Several other undesirable genetic pathways were detected in both
180 groups. For example, putrescine biosynthesis pathways and polymyxin resistance genes co-
181 occurred in 7/10 samples (Figure 4c), and these pathways were all attributed to *E. cloacae*, *E.*
182 *coli*, *K. pneumoniae*, or a combination of these three species. We detected several other
183 antibiotic resistance genes, including beta-lactamase genes and methicillin resistance genes,
184 in both groups (Figure S4). In addition, we found HGT-associated genes, including plasmid
185 maintenance genes and transposition genes, in both groups.

186

187 **Application of strain-level analysis to characterise enteric bacteria in nunu**

188 Leonard *et al.* previously used metagenomic sequencing to detect *E. coli* in spinach which
189 was intentionally spiked with *E. coli* O157:H7 strain Sakai (11). We downloaded the
190 metagenomic reads from that study (16 samples) and we subjected them to StrainPhlAn,
191 MetaMLST and PanPhlAn analysis, to confirm that these tools can accurately detect
192 pathogens in food samples: MetaMLST was used for multi-locus sequence typing,
193 StrainPhlAn was used for phylogenetic identification, and PanPhlAn was used for functional
194 characterisation. MetaMLST accurately detected *E. coli* ST11 in 7/16 spinach samples (Table
195 1). StrainPhlAn detected *E. coli* strains in 5/16 samples and it showed that the *E. coli* strain in
196 each of these samples was closely related to *E. coli* O157:H7 strain Sakai (Figure 5).
197 PanPhlan detected Shiga toxin genes in 15/16 samples (Table 1) and it indicated that the *E.*
198 *coli* strain in each of these samples was most closely related to *E. coli* O157:H7 strain Sakai.
199 Thus, overall, PanPhlAn was the most sensitive method in this instance, since it was able to
200 detect STEC in almost all of the samples, whereas the other tools detected STEC in less than
201 half of the samples. In a follow-on study, Leonard *et al.* spiked spinach with 12 different
202 Shiga toxin producing *E. coli* strains, and they detected single strains in 17 samples (18). We
203 downloaded the metagenomic reads from the 17 samples and ran PanPhlAn, and were able to
204 identify Shiga toxin genes in all 17 samples (Table S2).

205 Having established the relative merits of these tools, we subsequently employed all three
206 strategies to identify the strains of *E. coli* and *K. pneumoniae* present in the nunu samples.
207 With regard to *E. coli*, MetaMLST detected a novel *E. coli* sequence type in 1t7am (Table 2).
208 StrainPhlAn detected 24 *E. coli* marker genes in the samples and a phylogenetic tree (Figure
209 6a), which was generated by aligning these markers against 118 *E. coli* reference genomes

(listed in Table S3), revealed that the *E. coli* strain in one sample, 1t7am, was closely related to *E. coli* O139:H28 E24377A. PanPhlAn detected *E. coli* strains in two samples: 1t7am and 1t8am. MDS analysis indicated that the strains from the two samples were functionally distinct from one another. Notably, a ShET2 enterotoxin encoding gene was identified in the *E. coli* strain from 1t7am. The same gene was found in *E. coli* O139:H28 E24377A. With regard to *K. pneumoniae*, MetaMLST detected the known sequence type *K. pneumoniae* ST39 in the sample 2u3am. Apparently novel *K. pneumoniae* sequence types were identified in six other samples (Table 1). StrainPhlAn detected 38 *K. pneumoniae* marker genes in the samples and a phylogenetic tree (Figure 6b), which was constructed by aligning these markers against 40 *K. pneumoniae* reference genomes (listed in Table S4), revealed that the *K. pneumoniae* strains in two samples, 1t8am and 2u3am, were closely related to *K. pneumoniae* KpQ3. In contrast, the *K. pneumoniae* strain in1t7am was most closely related to *K. pneumoniae* UCICRE 7. MDS analysis of the PanPhlAn output showed that five of the detected *K. pneumoniae* strains were functionally similar to one another (Figure 6c). However, two of the detected *K. pneumoniae* strains, in samples 1t6am and 1t7am, appeared to be functionally distinct from the others. In addition, PanPhlan indicated that sample 1t6am might have contained multiple strains, since an unusually high number of 5746 *K. pneumonia* gene families were detected. A TEM beta-lactamase gene was found in 1t2am using PanPhlAn and, furthermore, an OXA-48 carbapenemase gene was detected in 2u8am and the same gene was found in *K. pneumoniae* KpQ3.

Finally, we compared the time taken to process 10 nunu metagenome samples using the short-read alignment tools versus the metagenome assembler IDBA-UD (Figure S5). In each case, we observed that all of the short-read alignment tools were faster than IDBA-UD. It is important to note that additional bioinformatics analyses (contig binning, SNP analysis, etc.) are required to achieve strain-level identification from assembled metagenomes, and this emphasises the superior speed of the short-read alignment tools.

## Discussion

Foodborne pathogens are responsible for millions of cases of disease annually, in the United States alone (43). High-throughput sequencing can potentially be used to detect pathogenic strains in food products to prevent the occurrence of disease outbreaks. A recent proof of concept study demonstrated that whole metagenome shotgun sequencing accurately detected

Shiga toxin producing *E. coli* (STEC) strains in spiked spinach samples (18). However, that study used whole metagenome assembly-based approaches to achieve strain-level taxonomic resolution of the STEC in the samples. Whole metagenome assembly is a computationally intensive, time-consuming process, as illustrated by Nurk *et al.*, who recently reported that metagenome assembly can take between 1.5 hours to 6 hours, with a memory footprint ranging from 7.3 GB to 234.5 GB, to process a single human gut metagenomic sample, depending on the chosen assembler (44). Thus, the application of more rapid, less intensive bioinformatic tools for strain detection is desirable. In this study, we demonstrate that the short read alignment-based programs MetaMLST, StrainPhlAn, and PanPhlAn can accurately identify pathogens in food products.

We validated the accuracy of each approach by processing spinach metagenome data from samples that were spiked with the STEC O157:H7 Sakai in a previous study (11). We observed that PanPhlAn was the most sensitive approach. Indeed, PanPhlAn was able to identify STEC in every sample where it was present at >2% relative abundance, whereas the other approaches worked best when STEC was present at high relative abundances. However, none of the tools detected *E. coli* O157:H7 Sakai in every sample tested. The observation of false negatives highlights that the tools are not entirely accurate. It is likely that increased sequencing depth and/or longer sequencing read lengths would reduce the false negative rate. We recommend that these tools be used to supplement data from metagenome sequence classifiers like MetaPhlAn2, which did detect *E. coli* in each sample. Therefore, we subsequently used the strain-level analysis tools in combination with other metagenomic approaches to assess the safety of nunu, a traditional Ghanaian fermented milk product.

Nunu is produced through the spontaneous fermentation of raw cow milk in calabashes or other containers for 24-36 hours at ambient temperature (23). The crude nature of the nunu production process has raised food safety concerns (25). Indeed, several potentially pathogenic microorganisms were previously detected in nunu samples by microbial culturing (25). This resulted in some nunu producers receiving hygiene practice training to improve food safety. However, our work suggests that there is little difference in the prevalence of pathogens in nunu samples from trained and untrained producers. One reason for this may be that it is difficult for the nunu producers to adhere to the training recommendations which are not appropriate to the rural production conditions. During training, the producers were advised to pasteurise the milk before cooling and adding a starter culture. After incubating for 4-6 hours in a covered container, they were advised to stir the mixture and refrigerate the

product. Lack of access to specific heat control and electricity, as well as the variance from the traditional method, which does not use a starter culture, are both reasons why the training is not adhered to.

16S rRNA gene sequencing revealed that the samples were dominated by Lactobacillales. However, we also detected high abundances of Enterobacteriales, including *Enterobacter* and *Escherichia,* in both groups. Subsequently, whole metagenome shotgun sequencing showed that most samples were dominated by *Streptococcus infantarius*, a species which was previously identified in other African dairy products (45, 46). Concernedly, *S. infantarius* has been linked to several human diseases, including bacteraemia (47), endocarditis (48) and colon cancer (49). Aside from *S. infantarius*, two other potentially pathogenic species, *Escherichia coli* and *Klebsiella pneumoniae*, were identified in a subset of samples.

Overall, our findings indicate that nunu samples from trained producers and untrained producers were contaminated with faecal material. Cattle faeces can be a major source of bacterial contaminants in raw cow milk (29), and thus, our results are not entirely surprising, but the remarkable abundance of such microorganisms in nunu is worrying. It had been hoped that nunu could be used to supplement traditional cereal-based weaning foods to improve infant nutrition. However, qualitative research among mothers and health workers highlighted safety concerns, which, as we have shown here, are valid. In particular, the presence of *E. coli* and *K. pneumoniae* in nunu is a concern, and, thus, we employed strain-level metagenomics for the further characterisation of these bacteria.

In terms of *E. coli*, strain-level analysis indicated that the *E. coli* strain in one sample was an enterotoxin producer and it was closely related to *E. coli* O139:H28 E24377A, a strain which was linked to an outbreak of waterborne diarrhoea in India (50). In terms of *K. pneumoniae,* strain-level analysis indicated that the *K. pneumoniae* strains in two samples were antibiotic resistant and they were closely related to *K. pneumoniae* KpQ3, a strain which was linked to nosocomial outbreaks among burn unit patients. Thus, strain-level analysis suggests that there are likely pathogens in some of the samples. Interestingly, PanPhlAn also suggested that there were functionally distinct strains of both species in nunu samples from different producers. Perhaps, this indicates multiple incidences or sources of contamination. Undoubtedly, our work highlights an urgent need to further improve hygiene practices during nunu production, and the pasteurisation of the starting milk and the use of starter-based fermentation systems is an obvious solution.

In conclusion, our work suggests that short read alignment-based strain detection tools can be used to detect pathogens in other foods, apart from nunu or spinach, and they might also be useful for tracing the sources of foodborne disease outbreaks back to particular foods. Such tools are a significant improvement over 16S rRNA gene sequencing, which is often limited to genus-level identification, or metagenome read classification tools, which are limited to species-level identification (16). In addition, they are faster, and less computationally intensive, than metagenome assembly-based strain detection methods, making them more relevant to real-life scenarios which necessitate the rapid testing of many food samples. With DNA sequencing costs continuing to decrease, the approach outlined here is an affordable option for food safety testing.

**Acknowledgements**

We would like to thank the researchers at the Animal Research Institute, in Accra, Ghana for their help with sample collection, storage and processing. This research was funded by Science Foundation Ireland in the form of a centre grant (APC Microbiome Institute grant number SFI/12/RC/2273). Research in the Cotter laboratory is also funded by Science Foundation Ireland through the PI award “Obesibiotics” (11/PI/1137). Orla O’Sullivan is funded by Science Foundation Ireland through a Starting Investigator Research Grant award (13/SIRG/2160). Kareem Daari is a PhD student funded by Ghana Educational Trust (GET) Fund. The Rowett Institute receives funding from the Scottish Government (RESAS).

## Materials and Methods

### Sampling

Five nunu samples were collected from producers with hygiene practice training, and another five samples were collected from producers without hygiene practice training. The identity of the samples from trained and untrained individuals was blinded until after sequencing analysis was completed. The samples from the trained group were labelled 1t2am, 1t6am, 1t7am, 1t8am, and 2t2am. The samples from the untrained group labelled 1u6am, 2u2am, 2u3am, 2u6am, and 2u8am. All samples were collected in the morning and placed on ice for transport to the lab. Sample aliquots (4ml) were then mixed with glycerol to a final concentration of 20% and stored at -20°C prior to DNA extraction. DNA was extracted from

the samples at the Animal Research Institute, Accra, Ghana and then sent to Scotland to comply with International laws on the import of animal samples (Import Licence form AB117).

**Microbiological analysis**

Basic microbiology culture analysis was carried out in Ghana. The plate-count technique was used to estimate the total viable bacterial count of the nunu samples on Milk Plate Count Agar (LAB M, UK). Bacterial counts were compared for plates growing aerobically or anaerobically at 30°C for 36-72 h. Anaerobic plates were incubated in airtight canisters containing $C0_2$Gen sachets (Oxoid, UK), which created an anaerobic atmosphere. Following incubation, colonies were counted using an SC6+ electronic colony counter (Stuart Scientific, UK). The presence of specific pathogens in the nunu samples was determined by streaking nunu directly onto selective agar plates to visually assess bacterial growth. The following selective agars were used: Blood agar (Merck, Germany) for *Staphylococcus*; MacConkey agar (Merck, Germany) for Enterobacteria; de Man Rogosa Sharpe agar (MRS) (Oxoid, UK) for *Lactobacillus* species; and *Salmonella Shigella* agar (Oxoid, UK). Any mixed growth plates were re-purified by streaking onto selected secondary agars. Lactose fermenting colonies identified on MacConkey agar were sub-cultured onto Eosin Methylene Blue Agar (EMBA) (Scharlau Chemie, Spain) to isolate/identify *E. coli.* Additionally, *Staphylococcus* colonies from Blood Agar were sub-cultured onto Mannitol Salt Agar (MSA) (Oxoid, UK) to isolate/identify *Staphylococcus aureus*. The following biochemical tests were used to confirm bacterial identification: the Motility Indole Urea (MIU) test; the catalase test; the Triple Sugar Iron (TSI) test; and the Indole Methyl Red Vorges-Proskeur Citrate (IMViC) tests. Cellular morphology was determined by Gram staining as well as microscopic examination.

**DNA extraction and next generation sequencing**

Briefly, 1 ml of each thawed sample was diluted in 9 ml of sterile PBS, mixed thoroughly using vortex and centrifuged for 10 min (8,000-10,000 g). The bacterial cell pellets were resuspended in 432 µl sterile $dH_2O$ and 48µl 0.5 M EDTA, mixed thoroughly by a combination of vortex and with a sterile pipette tip and the suspension frozen. The frozen samples were thawed on the bench and refrozen and finally thawed (giving a total of two

freeze/thaw cycles) before extracting the DNA using the Promega Wizard genomic DNA extraction kit (Promega, Madison, WI, USA) according to the manufacturer's protocol. The freeze/thaw cycles were carried out to maximise bacterial cell lysis. Following extraction, the DNA pellets were air dried for about 60 minutes and stored sealed under airtight conditions and transported from the Animal Research Institute, Accra, Ghana to the Rowett Institute, at University of Aberdeen, for further analysis.

DNA extracts were quantified using the Qubit High Sensitivity DNA assay (BioSciences, Dublin, Ireland). 16S rRNA gene sequencing libraries were prepared from extracted DNA using the 16S Metagenomic Sequencing Library Preparation protocol from Illumina, with minor modifications (26). Samples were sequenced on the Illumina MiSeq in the Teagasc sequencing facility, with a 2 x 250 cycle V2 kit, in accordance with standard Illumina sequencing protocols. Whole-metagenome shotgun libraries were prepared in accordance with the Nextera XT DNA Library Preparation Guide from Illumina (26). Samples were sequenced on the Illumina MiSeq in the Teagasc sequencing facility, with a 2 x 300 cycle V3 kit, in accordance with standard Illumina sequencing protocols.

**Bioinformatics**

Raw 16S rRNA gene sequencing reads were quality filtered using PRINSEQ (27). Denoising, OTU clustering, and chimera removal were done using USearch (v7-64bit) (28), as described by Doyle *et al.* (29). OTUs were aligned using PyNAST (30). Alpha-diversity and beta-diversity were calculated using Qiime (1.8.0) (31). Taxonomy was assigned using a BLAST search (32) against SILVA SSU 119 database (33).

Raw whole-metagenome shotgun sequencing reads were filtered, on the basis of quality and quantity, and trimmed to 200 bp, with a combination of Picard Tools (https://github.com/broadinstitute/picard) and SAMtools (34). MetaPhlAn2 was used to characterise the microbial composition of samples at the species-level (35). MetaMLST (20), PanPhlAn (19), and StrainPhlAn (21) were used to characterise the microbial composition of the samples at the strain-level. GraPhlAn (36) was used to construct phylogenetic trees from the StrainPhlAn output. SUPER-FOCUS (37) and HUMAnN2 (38) were used to determine the microbial metabolic potential of samples. IDBA-UD (39) was used for metagenome assembly.

## Accession numbers

Sequence data have been deposited in the European Nucleotide Archive (ENA) under the project accession number PRJEB20873.

## Statistical analysis

Statistical analysis was done in R-3.2.2 (40). The Kruskal-Wallis test was done using the compareGroups package, and the resulting p-values were for multiple comparisons. PCoA analysis of 16S rRNA gene sequencing data was done using the phyloseq package (41). Multidimensional scaling (MDS) was done using the vegan package. Data visualisation was done using the ggplot2 package.

## References


1. **Walsh AM, Crispie F, Claesson MJ, Cotter PD.** 2017. Translating Omics to Food Microbiology. Annual Review of Food Science and Technology **8**.
2. **Zheng J, Zhao X, Lin XB, Gänzle M.** 2015. Comparative genomics Lactobacillus reuteri from sourdough reveals adaptation of an intestinal symbiont to food fermentations. Scientific reports **5:**18234.
3. **Sun Z, Harris HM, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF.** 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nature communications **6**.
4. **Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W.** 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104: H4 outbreak by rapid next generation sequencing technology. PloS one **6:**e22751.
5. **Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, Petrovska L, Ellis RJ, Elson R, Underwood A.** 2015. Whole-Genome Sequencing for National Surveillance of Shiga Toxin–Producing Escherichia coli O157. Clinical Infectious Diseases**:**civ318.

6. **Kwong JC, Mercoulia K, Tomita T, Easton M, Li HY, Bulach DM, Stinear TP, Seemann T, Howden BP.** 2016. Prospective whole-genome sequencing enhances national surveillance of Listeria monocytogenes. Journal of clinical microbiology **54:**333-342.
7. **De Filippis F, Genovese A, Ferranti P, Gilbert JA, Ercolini D.** 2016. Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate. Scientific reports **6**.
8. **Quigley L, O'Sullivan DJ, Daly D, O'Sullivan O, Burdikova Z, Vana R, Beresford TP, Ross RP, Fitzgerald GF, McSweeney PLH, Giblin L, Sheehan JJ, Cotter PD.** 2016. Thermus and the Pink Discoloration Defect in Cheese. mSystems **1**.
9. **Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ, Cotter PD.** 2016. Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems **1:**e00052-00016.
10. **Yang X, Noyes NR, Doster E, Martin JN, Linke LM, Magnuson RJ, Yang H, Geornaras I, Woerner DR, Jones KL.** 2016. Use of Metagenomic Shotgun Sequencing Technology To Detect Foodborne Pathogens within the Microbiome of the Beef Production Chain. Applied and environmental microbiology **82:**2433-2443.
11. **Leonard SR, Mammel MK, Lacher DW, Elkins CA.** 2015. Application of metagenomic sequencing to food safety: detection of Shiga toxin-producing Escherichia coli on fresh bagged spinach. Applied and environmental microbiology **81:**8183-8191.
12. **Wang Q, Garrity GM, Tiedje JM, Cole JR.** 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology **73:**5261-5267.
13. **Allard G, Ryan FJ, Jeffery IB, Claesson MJ.** 2015. SPINGO: a rapid species-classifier for microbial amplicon sequences. BMC bioinformatics **16:**1.
14. **Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML.** 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods in Ecology and Evolution **4:**1111-1119.
15. **Stellato G, Utter DR, Voorhis A, De Angelis M, Eren AM, Ercolini D.** 2017. A few Pseudomonas oligotypes dominate in the meat and dairy processing environment. Frontiers in Microbiology **8**.
16. **Lindgreen S, Adair KL, Gardner PP.** 2016. An evaluation of the accuracy and speed of metagenome analysis tools. Scientific reports **6:**19233.

17. **Stasiewicz MJ, den Bakker HC, Wiedmann M.** 2015. Genomics tools in microbial food safety. Current Opinion in Food Science **4:**105-110.

18. **Leonard SR, Mammel MK, Lacher DW, Elkins CA.** 2016. Strain-Level Discrimination of Shiga Toxin-Producing Escherichia coli in Spinach Using Metagenomic Sequencing. PloS one **11:**e0167870.

19. **Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N.** 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature methods.

20. **Zolfo M, Tett A, Jousson O, Donati C, Segata N.** 2016. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Nucleic Acids Research**:**gkw837.

21. **Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, Ferretti P, Gorfer V, Pedrotti A, Tett A, Segata N.** 2017. Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling. mSystems **2**.

22. **Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, Weir JC, Quince C, Smith GP, Betley JR.** 2013. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104: H4. Jama **309:**1502-1510.

23. **Akabanda F, Owusu-Kwarteng J, Tano-Debrah K, Glover RL, Nielsen DS, Jespersen L.** 2013. Taxonomic and molecular characterization of lactic acid bacteria and yeasts in nunu, a Ghanaian fermented milk product. Food microbiology **34:**277-283.

24. **Marsh AJ, Hill C, Ross RP, Cotter PD.** 2014. Fermented beverages with health-promoting potential: past and future perspectives. Trends in Food Science & Technology **38:**113-124.

25. **Akabanda F, Owusu-Kwarteng J, Glover R, Tano-Debrah K.** 2010. Microbiological characteristics of Ghanaian traditional fermented milk product, Nunu. Nature and Science **8:**178-187.

26. **Clooney AG, Fouhy F, Sleator RD, O'Driscoll A, Stanton C, Cotter PD, Claesson MJ.** 2016. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. PloS one **11:**e0148028.

27. **Schmieder R, Edwards R.** 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics **27:**863-864.

28. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460-2461.

29. **Doyle CJ, Gleeson D, O'Toole PW, Cotter PD.** 2017. Impacts of Seasonal Housing and Teat Preparation on Raw Milk Microbiota: a High-Throughput Sequencing Study. Applied and Environmental Microbiology **83:**e02694-02616.

30. **Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R.** 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics **26:**266-267.

31. **Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nature methods **7:**335-336.

32. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. Journal of molecular biology **215:**403-410.

33. **Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO.** 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research **41:**D590-D596.

34. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.** 2009. The sequence alignment/map format and SAMtools. Bioinformatics **25:**2078-2079.

35. **Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N.** 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature methods **12:**902-903.

36. **Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N.** 2015. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ **3:**e1029.

37. **Silva GGZ, Green KT, Dutilh BE, Edwards RA.** 2016. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics **32:**354-361.

38. **Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B.** 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS computational biology **8:**e1002358.

39. **Peng Y, Leung HC, Yiu S-M, Chin FY.** 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28:**1420-1428.

40. **Team RC.** 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. ISBN 3-900051-07-0.
41. **McMurdie PJ, Holmes S.** 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one **8:**e61217.
42. **Fukushima M, Kakinuma K, Kawaguchi R.** 2002. Phylogenetic analysis of Salmonella, Shigella, and Escherichia coli strains on the basis of the gyrB gene sequence. Journal of clinical microbiology **40:**2779-2785.
43. **Scallan E, Hoekstra R, Mahon B, Jones T, Griffin P.** 2015. An assessment of the human health impact of seven leading foodborne pathogens in the United States using disability adjusted life years. Epidemiology and infection **143:**2795-2804.
44. **Nurk S, Meleshko D, Korobeynikov A, Pevzner PA.** 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Research**:**gr. 213959.213116.
45. **Abdelgadir W, Nielsen DS, Hamad S, Jakobsen M.** 2008. A traditional Sudanese fermented camel's milk product, Gariss, as a habitat of Streptococcus infantarius subsp. infantarius. International journal of food microbiology **127:**215-219.
46. **Jans C, Kaindi DWM, Böck D, Njage PMK, Kouamé-Sina SM, Bonfoh B, Lacroix C, Meile L.** 2013. Prevalence and comparison of Streptococcus infantarius subsp. infantarius and Streptococcus gallolyticus subsp. macedonicus in raw and fermented dairy products from East and West Africa. International journal of food microbiology **167:**186-195.
47. **Beck M, Frodl R, Funke G.** 2008. Comprehensive study of strains previously designated Streptococcus bovis consecutively isolated from human blood cultures and emended description of Streptococcus gallolyticus and Streptococcus infantarius subsp. coli. Journal of clinical microbiology **46:**2966-2972.
48. **Herrero IA, Rouse MS, Piper KE, Alyaseen SA, Steckelberg JM, Patel R.** 2002. Reevaluation of Streptococcus bovis endocarditis cases from 1975 to 1985 by 16S ribosomal DNA sequence analysis. Journal of clinical microbiology **40:**3848-3850.
49. **Biarc J, Nguyen IS, Pini A, Gossé F, Richert S, Thiersé D, Van Dorsselaer A, Leize-Wagner E, Raul F, Klein J-P.** 2004. Carcinogenic properties of proteins with pro-inflammatory activity from Streptococcus infantarius (formerly S. bovis). Carcinogenesis **25:**1477-1484.
50. **Tamhankar AJ, Nerkar SS, Khadake PP, Akolkar DB, Apurwa SR, Deshpande U, Khedkar SU, Stålsby-Lundborg C.** 2015. Draft genome sequence of

566 enterotoxigenic Escherichia coli strain E24377A, obtained from a tribal drinking
567 water source in India. Genome announcements **3:**e00225-00215.

568

569 **Table 1. The results of MetaMLST and PanPhlAn analysis of spinach metagenomes**
570 **spiked with *E. coli* O157:H7 Sakai**

| Sequence accession number | Reads | *E. coli* abundance (%) | stx2A | stx2B | Sequence type (ST) | Confidenc (%) |
|---|---|---|---|---|---|---|
| SRR2177250 | 9,365,812 | 5.28412 | 1 | 1 | Unknown | NA |
| SRR2177251 | 17,562,542 | 4.31712 | 1 | 1 | 11 | 99.97 |
| SRR2177280 | 11,707,292 | 21.16364 | 1 | 1 | 100001 | 99.97 |
| SRR2177281 | 10,580,532 | 2.84187 | 1 | 1 | Unknown | NA |
| SRR2177282 | 6,155,636 | 60.51406 | 1 | 1 | 11 | 100 |
| SRR2177283 | 13,120,244 | 10.11327 | 1 | 1 | 11 | 100 |
| SRR2177284 | 7,500,056 | 2.05064 | NA | NA | Unknown | NA |
| SRR2177285 | 14,482,370 | 66.69813 | 1 | 1 | 11 | 100 |
| SRR2177286 | 14,035,970 | 69.17834 | 1 | 1 | 11 | 100 |
| SRR2177287 | 12,242,348 | 5.62746 | 1 | 1 | Unknown | NA |
| SRR2177288 | 8,303,788 | 10.75005 | 1 | 1 | 11 | 100 |
| SRR2177357 | 14,621,672 | 8.02047 | 1 | 1 | 11 | 100 |
| SRR2177358 | 10,684,052 | 3.18652 | 1 | 1 | Unknown | NA |
| SRR2177359 | 4,964,436 | 1.17146 | 1 | 1 | Unknown | NA |
| SRR2177360 | 12,729,834 | 1.81229 | 1 | 0 | Unknown | NA |
| SRR2177361 | 11,946,092 | 0.70921 | 0 | 1 | Unknown | NA |

571

572 **Table 2. The results of MetaMLST analysis of the nunu metagenomic samples**

| Species | Sequence type (ST) | Confidence (%) | Sample |
|---|---|---|---|
| *Klebsiella pneumoniae* | 100001 | 98.7 | 1t2am |
| *Klebsiella pneumoniae* | 100002 | 100 | 1t6am |
| *Esherichia coli* | 100001 | 100 | 1t7am |
| *Klebsiella pneumoniae* | 100003 | 99.9 | 1t7am |
| *Klebsiella pneumoniae* | 100004 | 100 | 1t8am |
| *Klebsiella pneumoniae* | 39 | 100 | 2u3am |
| *Klebsiella pneumoniae* | 100005 | 99.91 | 2u6am |
| *Klebsiella pneumoniae* | 100006 | 99.91 | 2u8am |

573

574

575 **Figure legends**

576 **Figure 1. 16S rRNA gene sequencing based analysis of nunu samples.** (A) Heat map
577 showing the 25 most abundant bacterial genera across the nunu samples. (B) Bar plot shoing
578 genera which were differentially abundant in either group.

579 **Figure 2. The species-level microbial composition of nunu samples, as determined by**
580 **MetaPhlAn2.**

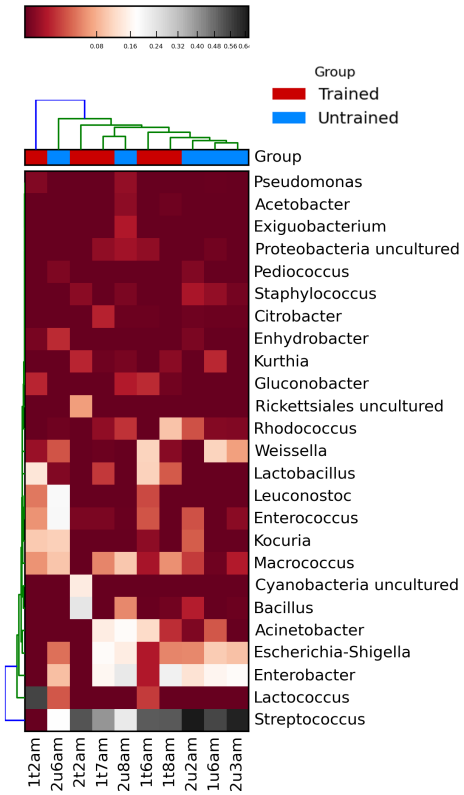
581 **Figure 3. The average abundances of the SUPER-FOCUS Level 1 functions that were**
582 **detected in nunu samples.**

583 **Figure 4. HUMAnN2 analysis.** (A) Heat map showing the 25 most abundant MetaCyc
584 pathways detected across the ten nunu metagenomic samples. (B) Bar plot showing
585 differences in histidine metabolic potential between nunu samples from trained producers and
586 nunu samples from untrained producers. (C) Bar plots showing the relative contributions of
587 *E. cloacae*, *E. coli* and *K. pneumoniae* to the MetaCyc pathways PWY-6305 (putrescine
588 biosynthesis) and PWY0-1338 (polymyxin resistance).

589 **Figure 5. StrainPhlAn analysis of the spinach metagenome.**

590 **Figure 6. Strain-level analysis.** Phylogenetic trees showing the relationships between (A) *E.*
591 *coli* strains and (B) *K. pneumoniae* strains detected in the nunu metagenomic samples and
592 their respective reference genomes, as predicted by StrainPhlAn. (C) MDS showing the
593 functional similarities between strains detected in the nunu metagenomic samples, as
594 predicted by PanPhlAn; reference genomes are shown in faded grey.

**a**




**b**